Supplementary Table 1: FISH probes

Gli1 Intron	Gli1 Exon1	Gli1 Exon2	RG6 Intron
AATCTAGGGAGGGATGGGT	AGACGGCGAGACACAGGTG	CAATCCGGTGGAGTCAGAC	GAGGTGGAGAGATGGAACAA
AACTAGGACCCTACCTTGA	GGCTGACTGTGTAAGCAGA	TTCACTGGGGTGGGCATTG	GACCCGCGATTTATTCACAG
GGAAGGAACTTTGAGGCCA	GAAAGGGGATGCCAGGGAG	GAAAACAGAGGCTGCGGGC	AGGGAAAGGCACAGGACACA
GTCGATACAGTCTTCAGCG	AGGCCAGGTAGTGACGATG	TGAGGGAGCTGGGGATGAT	TGAAAGGAAAGCGTTTCCCA
CCACCCCACAAAAATGCAA	ATGTAGTGCTGAGCAGGTG	GACCCGACTGGGGATACTG	GGATGGGGAGAGCGCAAAAA
CCATATACACACCTTGGGA	GAAGCATATCTGGCACGGA	CATGGGGAGGCTGAGGATA	CAGAACTGAGAGCAGGTTGG
CTGAGAGTTGTCCTCTGAC	CAGGCTGTGAGCTGCAGTG	CTGGGTTCTGTTGAGAGAT	GTGTTGGGATGGAAGACAGA
TGCTTTTGATCTCCAGGAC	ACAGAAAGACCTCCCATCC	AGGGGTTGAAATTGAGGCC	AAGAAGACAGCAACCCAGGG
AAGTGTCTTTAGCCAAGCC	GTTCGGCTTCTCCAAGGAG	GTCCTGTGGAATGAGAGGA	TCCCACAGCAAGAGAGGAAG
CATTTTAGGGCCAGCAAGA	CTGCATTTGGGTTGTATCC	ACACCAGCTGAGCTTTGAG	ACGAGAGTCCTCTGTCAGAG
TGTGTACCACCATACCAAA	ACATCCCAGGCTCTTGAAC	GCTGCGACTGAACGTAATT	ACACTGCTGGAGCATTTTTC
CTTCACAAGTGCTAGGGTT	GTGGTGGGGATCGAAGTTC	TTCCCTCCCACAACAATTC	TATGGCATAAATCTCCTGCA
CTCTTGATCTTCCTGACTC	GGCTGTGGCGAATAGACAG	ATGGGAGTTCCTGGTTGGG	CCATGCAGCAGAGAGCAAAA
TTTTTTGAGACCTAGCCTC	TCCATGGCAACATTTTCGG	CCAGAAACTTGGGGCTCTG	AGAGTGGGGGAGATCGACTT
TTCTAGCACCTTGCTTTTG	CCCATCACAGAAGTTCCAA	GGCTCTGACTAACTTGGGA	TAAATCCAAGGACTCGGAGC
GAACACAATGCCACTGACC	CATGTATGGGTTCAGACCA	AAGCCAGATCCATATGCTG	TACCTGAAGGTAAGGCATGG
GCTGGGGATTGAACTCAGA	CCAGAGTATCAGTGGAGGA	CTGATTTGTGATTGGCCGA	AGTAAAGGGCTCTGTTCAGG
GGTCATGAGCTAGCATGTA	TAGCTTCATAAGGCTCAGC	GTGAAGGGGCAGGATAGGA	GCTATGGGAGAAAGAAGCAG
GACAGAAGAGGGCAGCAGA	AAGAGGCAGGGAACCTGGA	CCCACGGTGAAAGTTTCAT	TCCAGGAGACCCGAAGAGAG
ATGAGTGCTCTATCTGCAT	CCATAGTTGGTTGGTGGAC	CTGTGGGAAGGCCTGTTTA	AGAATGGACAACCGTGGCAC
CCTGGTTTGGTTTTTGTTT	GGATCAGGATAGGAGACCT	AGAAGTCGGGGTGGTGCTG	TAAGGACAGGGTCAGTAGGA
TGGGATTAAGGGTGTGTGC	CCCAGCATGAGAAGGGAAC	CATAGCAAGGGGACAGCGG	
AAGAGTTCTGACTGCCTGT	AGCCTTATTGCTAGGGTAC	CACAGCTGGGGTTGGTATC	
GACCAGACTGGCCTAAGAG	GTCGAGGACACTGGCTATA	TCAGGAGGAGGGTACAAGG	
TATCCTGGAGCTCATCATG	GCACTTGTCCATAATGCTC	AGTCCAGAGCGTTACACAC	
CTTATCAGATCGAGCCTCA	ACCCTTGTTCTGGTTTTAC	GAGTGTTGTCCAGGTCAAG	
TCAGAGGGGGAAATATGCT		CTCATCTAGGATAGCCACA	
ACATGGAGAGACCTTGTCT		CATGGGAAAGAGGAGGGCT	
ACACAGCAAGTTCCAGGAC		AGGGAGATGGGGTGTTTTT	
CTTGAGTTCAAAGCCAGCT		AGACACTCATGTTACCCAC	
ACACCTTTGATCACAGTGC		TGTCTCTCCAGGCAGAGAC	
AAAGGGACTGGGTAGTGGT		TAGGCACTAGAGTTGAGGA	
CCCAGCACATGTTTTCATT			
CTTCCAAATGCTGGGGTTA			
CACAGAGATATGCTTGCCT			
CACTATGTTAGACCAGGCT			
AGTTTTGGCTAGCCTTGAA			
TTTCTTGGAGACAGGGTCT			
TCTATCCACTAGGCAATGA			
TCAGGCTTACACTTGTGTC			
CAAGAGTGGGGTCATCTGG			
CTGCACAGGGCTTAGATGA			
GTACTGAGCATCTTGGAGC			
AAACAGCGCAAGGGGAGGG			
GCTAAAGGCAGAGGAAGCC			
GGAGAATCCCAGGATTAGG			
ATGGGAGAACATGGCGACC			
CAGACGGGACGTGGAGATT			

Supplementary Table 2: qPCR primers probes

	Forward primer	Reverse primer
RG6 unspliced isoform	TACAAGGATGACGATGACAAGG	TGAACCAAAGCAGCAGGAG
RG6 unspliced isoform	CAGCCACACATCCTGAGAGC	AGCAGAGGTGGAGAGATGGA
RG6 spliced isoform	TCCGGAAGAATTCAGGTCAGGAG	GCGCATGAACTCCTTGATGAC
RG6 spliced isoform	TACCCGGATCTAGAGGTCAGGAG	GCGCATGAACTCCTTGATGAC

Supplementary Note 1: Quantification of smFISH dot intensity

Image acquisition

We used confocal microscopy to image FISH probes in Z-stacks with a step size of 0.2um. This step size is small enough to capture the maximum intensity of the FISH dot (Figure S1b). As demonstrated previously, the Z-maximum projection is comparable to 3D fitting[1]. In the following steps, we will use Z-maximum projection, instead of 3D fitting, to simplify the quantification of dot intensity.

Dot identification

We first performed a 2D Gaussian wavelet transform on the original image, a step comparable to the conventional filtering protocol in most FISH dot-counting methods[2]. Specifically, the transformed image equals:

$$H_{i0,j0} = \sum_{i,j} \left(D_{i,j} - B_{i,j}\right) \mathrm{G}_{i-i0,j-j0;\sigma}$$

where $D_{i,j}$ is the original intensity at pixel (i, j); $\mathrm{G}_{i-i0,j-j0;\sigma}$ is a truncated 2D Gaussian filter, defined as $\mathrm{G}_{\Delta i,\Delta j;\sigma} = \frac{1}{2\pi\sigma^2} e^{-\frac{\Delta i^2+\Delta j^2}{2\sigma^2}}$, $\Delta i, \Delta j \leq 7$; and $\mathrm{B}_{i,j}$ is the local background around pixel (i, j), defined as $\mathrm{B}_{i,j} = \sum_{p,q=-7}^{7} (D_{i-p,j-q})(\sum G)/n$, where $\sum G \cong 1$, $n = 225$ (i.e. the area covered by the truncated 2D Gaussian filter), and $\sigma \cong 1$ (i.e. the approximate value of the Gaussian standard deviation of a real smFISH dot). This step selected all potential FISH dots in the local area, including 'real' FISH dots as well as background-level dots.

FISH dot intensity fitting

After identifying the location of dots in each channel by wavelet transform, we chose a window centered on the selected dots in the original image, and then fit the raw fluorescent intensity. The fitting process was adapted from astrophysics for estimating stellar luminosities in crowded star fields[3], specifically using asymmetric 2D Gaussian integral with angle (θ):

$$\iint A \left(e^{-\left(\frac{\cos^2(\theta)}{2\sigma_x^2}+\frac{\sin^2(\theta)}{2\sigma_y^2}\right)(x-cx)^2+\left(\frac{\sin(2\theta)}{2\sigma_x^2}-\frac{\sin(2\theta)}{2\sigma_y^2}\right)(x-cx)(y-cy)-\left(\frac{\sin^2(\theta)}{2\sigma_x^2}+\frac{\cos^2(\theta)}{2\sigma_y^2}\right)(y-cy)^2} \right) dx\, dy$$

To deal with crowding of fluorescent dots in the image, we implemented a method used in stellar photometry of crowded star fields[3]. This method has two stages. In the first stage, one iteratively fits an image containing stars or, in this case, dots with a 2D Gaussian intensity distribution, removes it from the image, and then repeats this process for the next star or dot. This continues until the intensity of the putative dot falls below a threshold (here, 10% of the first integrated dot intensity), producing a set of *m* possibly overlapping Gaussian objects. In the second stage, one re-fits the original image to a linear combination of *m* Gaussians, whose positions are constrained to be close to the positions identified in the first stage (Figure S1c).

Supplementary Note 2: Quantifying intensity unit in each fluorescent channel

Fitting the intensity unit by a continuous analog of Poisson distribution

We fit the histogram of dot intensity with a continuous Poisson curve:

$$P = \frac{P_0 \left(\frac{x}{x_0}\right)^{\lambda} e^{-\lambda}}{\Gamma\left(\frac{x}{x_0} + 1\right)}$$

Here, P is the probability of distribution (i.e. y-axis of the histogram), and x is the the intensity of dots (i.e. x-axis of the histogram). The x value at the maximum P is the intensity unit (Figure 3b and Figure S1d). Note that the histogram did not explicitly exclude the TAS. However, because they represent only one dot out of thousands, in the tail of the distribution (Figure 3b and S1d), they do not affect the calibration significantly.

Obtaining the unbiased distribution of dot intensity in three channels

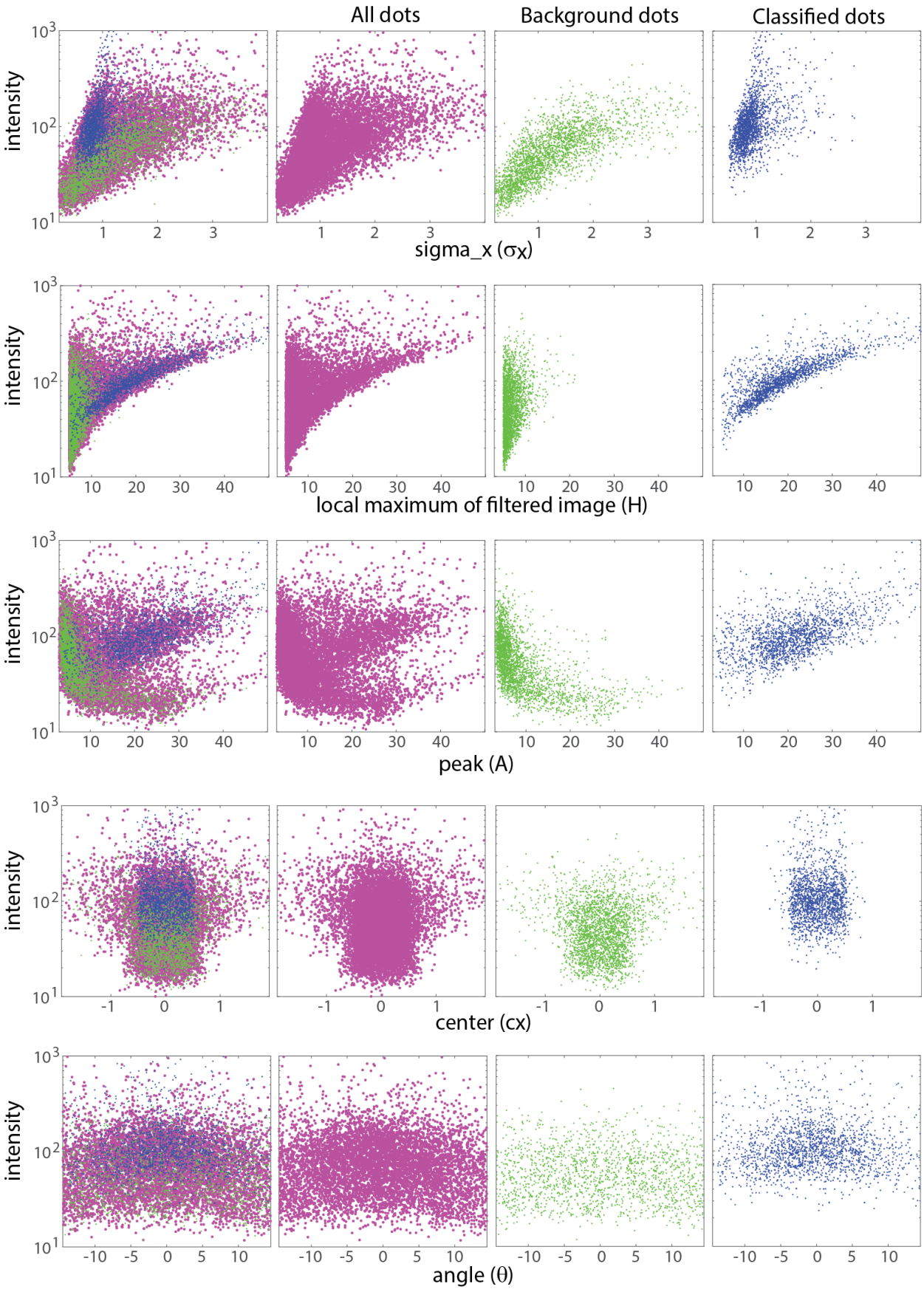
As described in the main text, we included some background-level dots in each channel to avoid dot-identification bias among different channels. For the background-subtraction method, we used the 'fmincon' function in Matlab to find the best Poisson fit parameters for the background-subtracted histogram (Figure 3a). To characterize the error of this method, we performed >100 fittings with varying sizes of the histogram bin and randomly picked subsets of the dot-intensity data to create histograms. As shown in Figure 3c, the variance of the obtained intensity units was less than 10% in all three channels. For the dot-colocalization method, despite the involvement of background dots, the probability of misclassifying a true background dot as a transcript was low, because misclassification requires at least co-localized background dot in another channel.

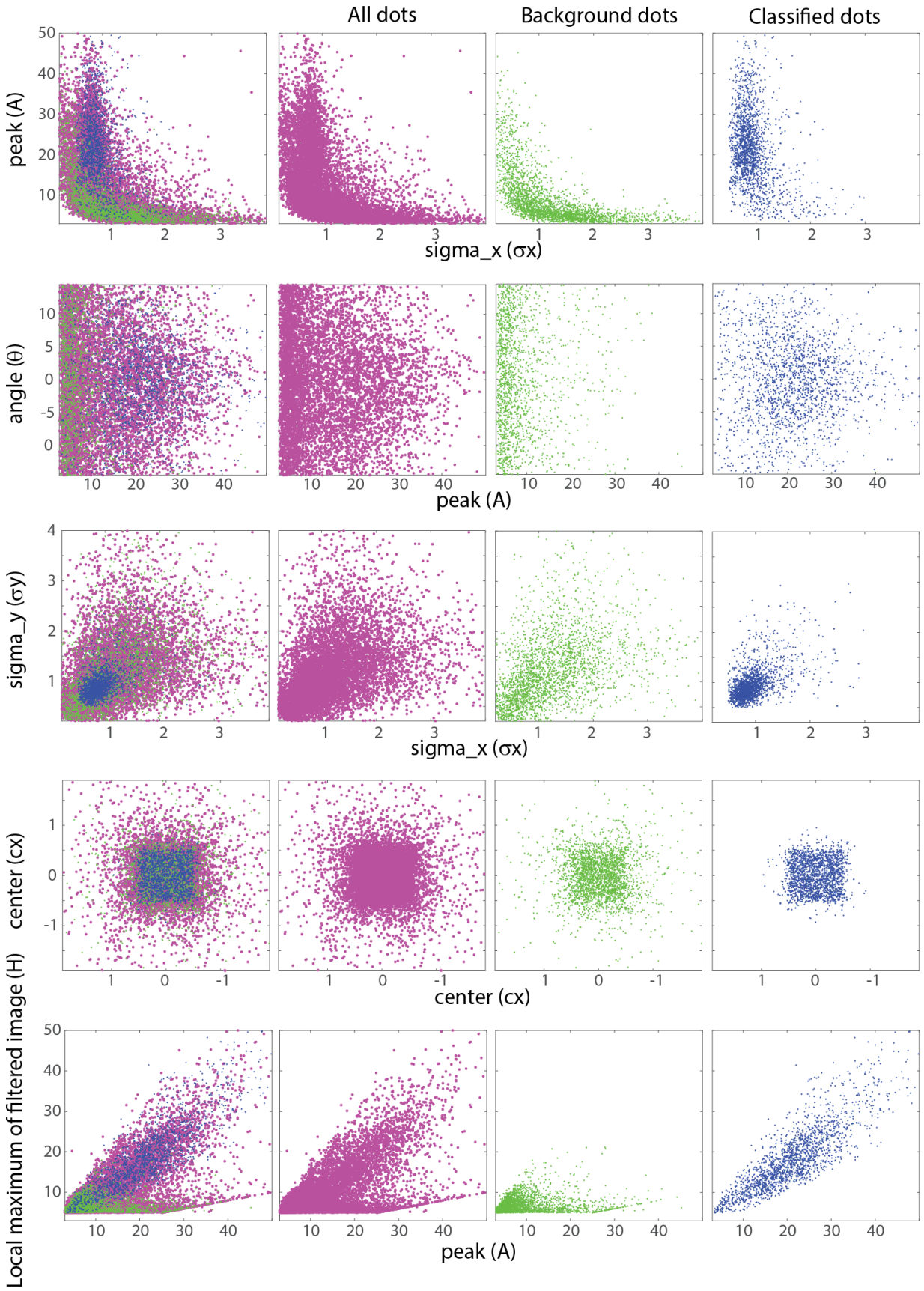
Picking background dots

The number of included background-level dots should not be substantially beyond the number of visible FISH dots. Otherwise, the histograms of background+foreground dots would be very similar to the histograms of background dots, making it impossible to perform the background-subtraction method. In addition, the chance to misclassify co-localized background dots to RNA transcripts would become high, distorting the measured intensity unit. In this paper, the number of involved background dots was comparable to (or less than) the number of visible FISH dots in each channel.

The fitting properties of background dots and classified dots

As shown in the figures below, the intensities of background dots were in general lower than the intensities of classified dots, but with some exceptions. These high-intensity background dots were due to two types of 'bad' dots: hot pixels (high fitting peak [A] with low sigma [σ_x, σ_y]), and dim speckle (low A with high σ_x, σ_y). This could be seen by the anti-correlation between fitting peak (A) and sigma (σ_x, σ_y) for background dots. In contrast to background dots, for foreground dots, sigma (σ_x, σ_y) was independent from peak (A). Two more differences existed when comparing fitting properties of foreground dots to background dots: (1) the distribution of angle (θ), σ_x versus σ_y, and center cx versus cy of classified dots


All dots
Background dots
Classified dots
intensity
sigma_x (σx)
local maximum of filtered image (H)
peak (A)
center (cx)
angle (θ)


All dots
Background dots
Classified dots
peak (A)
sigma_x (σx)
angle (θ)
peak (A)
sigma_y (σy)
sigma_x (σx)
center (cx)
center (cx)
Local maximum of filtered image (H)
peak (A)

were smaller, and (2) the fitting peak (A) of classified dots was more comparable to its local maximum of filtered image (H). These features of fitted dots indicate that true smFISH dots fall into a more reasonable ranges of 2D Gaussian fit parameters.

Comparison to conventional FISH quantification

As described in the main text, the background dots and classified dots had overlapping distributions on the values of local maximum of filtered image (H) and fitting peak (A) and would therefore be indistinguishable via conventional FISH protocol, which involves setting a threshold on H to identify FISH dots. Note that the other fit parameters, sigma (σ_x, σ_y), center (cx, cy) and angle (θ), were also overlapping. Consequently, setting parameter thresholds also cannot reduce rate of identifying false positive dots. Thus, our method is crucial to obtaining unbiased intensity unit in multiple channels.

Supplementary Note 3: 'Economy of scale' measurements

Data heterogeneity: the presence of 'negative' splicing efficiency

In principle, splicing efficiency ($1 - N_I/N_{E1}$) should always be positive, because the number of total transcripts N_{E1} cannot be smaller than the number of pre-spliced transcripts N_I. However, when the number of transcripts was quantified only based on the intensity of bound probes, the exact number of bound Exon1 probes could be less than the number of bound Intron probes at some transcription active sites (TASs) due to the stochastic binding of smFISH probes. Thus, for these TASs, measured N_{E1} could be smaller than N_I, resulting in an observed data points with 'negative' splicing efficiency and increasing the heterogeneity of splicing efficiency calculation.

Data heterogeneity: transcriptional bursting

Apart from the noise of experimental measurements, intrinsic 'transcriptional bursting' and extrinsic noise upstream of the transcription are also responsible for the cell-to-cell differences. Previous work has established a general model in which gene expression occurs through stochastic bursts and quantitative expression level distributions can be used to infer the burst rate and mean burst size[2]. As shown in Figure S7, the induction of our promoter (Tet-on CMV) primarily affects burst size, which are similar from those reported previously.

Data heterogeneity: geometric mean versus arithmetic mean

Apart from the stochastic binding of smFISH probes and transcriptional bursting, two other aspects could be responsible for the heterogeneity of the data points in Figure 4 and Figure S2: cell-cell variability and the static measurement (in fixed cells) of instantaneous splicing efficiency. To average out these sources of stochastic noise, we compute the geometric mean, because taking the arithmetic mean distorts the calculation and can even generate false-positive 'economy of scale' observations. We illustrate this point via a proof-of-concept example below. Set the true number of Intron and Exon1 value N_I, N_{E1} =2. If the measured N_I, N_{E1} = {1,2,3} (due to noise), the arithmetic mean of N_I/N_{E1}, i.e. (1/1 + 1/2 + 1/3 + 2/1 + 2/2 + 2/3 + 3/1 + 3/2 + 3/3)/9, equals 1.22, and the geometric mean equals 1. If the measured N_I, N_{E1} = {0.5, 1, 1.5, 2, 2.5, 3}, the arithmetic mean increases to 1.43, and the geometric mean still

equals 1. The change of arithmetic mean is due to the fractional nature of splicing efficiency: the denominator is more sensitive to stochastic noise at lower levels. Using the geometric mean helped eliminate the uneven effect of noise in the denominator and numerator values on the calculation of splicing efficiency.

TAS classification

We classified TASs based on co-localization of smFISH dots and the brighter dot intensity (Figure 2a). However, our method cannot perfectly distinguish low-expression TASs (i.e. dimmer smFISH dots) from dispersed unspliced transcripts. Here we set a specific condition to classify TAS: either $N_i > 2.5$, or $N_I + N_{E1} + N_{E2} > 5$. We chose the threshold based on the intensity distribution of single transcripts (Figure 3b and Figure S1d). In addition, we discarded the TASs with many emerging single transcripts, because this condition reflects different residence times of isoforms at the TAS, which is difficult to resolve in our protocol (discussed further in Supplementary Note 4).

DNA-FISH

We validated our observation of 'economy of scale' regulation by combining RNA- and DNA-FISH. False positives could be observed when misclassifying unspliced transcripts as TASs, because the splicing efficiency of any unspliced transcripts should always be zero, distorting the curve towards zero at low-transcription levels. To rule out this possibility, we repeated our experiments in combination with DNA-FISH. Specifically, we first performed RNA-FISH as previously described, and then did DNA-FISH in the same cells (SI Methods and Materials). We used the DAPI channel to register these images, and identified TASs by co-localization of both RNA- and DNA-FISH dots. Although fewer data points were acquired due to the complexity of these experiments, we observed the same 'economy of scale' effect (Figure S3).

'Hardness ratio' correction

Due to the spurious correlation between N_{E1} and $(1-N_I/N_{E1})$, using a single measurement of transcription level N_{E1} produces a false-positive 'economy of scale' observation (Figure S4a). A similar effect occurs for the control measurement (N_{E2} versus $(1-N_{E1}/N_{E2})$) as well. As described in the main text, we can eliminate the effect by using two independent transcription-level read-outs. Another possible solution is to use 'hardness of ratio' correction methods[4,5]. Converting N_I/N_{E1} to $N_I/N_{E1}(1+1/aN_{E1}+2/a^2N^2_{E1})$, and N_{E1}/N_{E2} to $N_{E1}/N_{E2}(1+1/aN_{E2}+2/a^2N^2_{E2})$ corrects the false-positive 'economy of scale' observation (Figure S4b).

Supplementary Note 4: A mechanism for 'economy of scale'

A phenomenological model for 'economy of scale'

As described in the main text, we proposed a model of non-uniform enzyme accessibility to explain the 'economy of sale' observation, where non-uniform enzyme accessibility represents a non-linear cooperativity between transcription and splicing factor recruitment. As shown in Figure S6a, when pre-mRNAs have a uniform enzyme accessibility (i.e. constant k_{on}) in the Michaelis-Menten model, the splicing efficiency should be close to 1 at low transcription levels

due to sufficient available enzymes and should only decrease at very high transcription levels due to enzyme titration in the system. In contrast, the non-uniform enzyme-accessibility model gives rise to the 'economy of scale' effect. (For simplicity, we consider here a model in which k_{on} is proportional to S, the available pre-mRNA, but other forms will have qualitatively similar results.) The splicing efficiency is close to zero at low expression levels, because there are too few pre-mRNAs to recruit sufficient splicing enzymes. When transcription level increases, the enzyme accessibility, and thus the splicing efficiency also increases, generating the 'economy of scale' behavior. Though the exact mechanism of this non-linear cooperativity remains unclear, analyzing protein liquid-liquid phase separation is a possible direction to pursue. Of note, these two models only differ at low transcription levels. At very high transcription levels, the curves overlap due to the enzyme titration effect in the system (Figure 5b and Figure S6a). Within our experimental system, we have not observed this titration effect by overexpressing a single target gene. This indicates that the physiological transcription level is not sufficient to titrate the splicing machinery in the cell.

Simulation of the Michaelis-Menten model

Based on Figure 5a, we have:

$$\begin{cases} E_{ext} + E + ES = E_0 \\ \frac{d}{dt}E_{ext} = -D_{in} \cdot E_{ext} + D_{out} \cdot ES \\ \frac{d}{dt}E = -k_{on} \cdot E \cdot S + k_m \cdot ES + k_{off} \cdot ES \\ \frac{d}{dt}S = b - k_{on} \cdot E \cdot S + k_{off} \cdot ES - k_u \cdot S \\ \frac{d}{dt}m = k_m \cdot ES - g_m \cdot m \\ \frac{d}{dt}u = k_u \cdot S - g_u \cdot u \end{cases}$$

The total level of enzymes (i.e. splicing factors) $E_0 = E + ES + E_{ext}$ remains constant in the cell, due to auto-regulation of splicing factors[6]. D_{out} and D_{in} are the diffusion rates of splicing factors 'out' and 'in' from the TAS, b is the transcription rate. S is the substrate (i.e. pre-mRNA). m is the mRNA (i.e. spliced isoform), u is the unspliced isoform, k_{on} and k_{off} are the binding and unbinding rates of splicing factors, k_m is the production rate of mRNA, g_u and g_m are the degradation rates of unspliced isoform and mRNA respectively.

At steady state, we set the time derivatives to zero and have:

$$\begin{cases} b - k_m \cdot ES - g_u \cdot S = 0 \\ K \cdot S \cdot \frac{E_0 - ES}{1 + rD} - ES = 0 \\ k_m \cdot ES - g_m \cdot m = 0 \end{cases}$$

where $rD = \frac{D_{out}}{D_{in}}$ and $K = \frac{k_{on}}{k_m + k_{off}}$. The splicing efficiency at the TAS is represented by $\frac{m}{S+ES+u+m} = \frac{\frac{b-g_u S}{g_m}}{S+\frac{b-g_u S}{k_m}+\frac{k_u}{g_u}S+\frac{b-g_u S}{g_m}}$. Figure 5a was obtained by setting $k_u = 0.1$, $k_m = 10$, $g_u = 0.1$, $g_m = 0.1$, $rD = 10$, $E_0 = 1500$, and $K = 0.5$. We modeled non-uniform enzyme accessibility by modifying $K = \frac{k_{on}}{k_m + k_{off}}$. To $\frac{k_{on}}{k_m + k_{off}} \cdot S$, where $\frac{k_{on}}{k_m + k_{off}}$ is a constant, achieving 'economy of scale' behavior (Figure 5b and S6a).

The observation of 'economy of scale' and 'diminishing returns' is not sensitive to parameters

The difference between 'economy of scale' and 'diminishing returns' can be represented by the 'sign' of the curve, i.e. splicing efficiency increasing (i.e. positive slope) or decreasing (i.e. negative slope) with transcription level. To simply the simulation, we defined the 'sign' as the difference of splicing efficiency in between b=1 and b=10. As shown in Figure S6c, scanning parameter-values only changes the magnitude of the splicing kinetic slope, but not the 'sign'.

Measuring the distance between a TAS and its nearest speckle

As shown in Figure 5C, we targeted Intron and Exon2 regions of RG6 and quantified the splicing efficiency as described in the previous section. We then performed immunostaining of splicing factor SC35 in the same cell (SI Materials and Methods), and measured the distance between each TAS and its nearest speckle (Figure S5a). Specifically, we first normalized the intensity of SC35 in each cell, because the total enzyme level should be constant based on the auto-regulation of splicing factors[6], and set an intensity threshold to determine the presence of speckles. We then measured the maximum SC35 intensity within a fixed distance of the TAS (in unit of pixels). Finally, we identified the distance at which the maximum intensity reached the intensity threshold. We then obtained the real distance between the TAS and its nearest speckle by multiplying by pixel size (129nm = 1pixel).

Other possible mechanisms for the 'economy of scale' observation

Modifying the enzyme-binding rate from k_{on} to $k_{on}S$ is not the only way to represent non-uniform enzyme accessibility. In principle, we can also tune the enzyme-diffusion parameter rD, replacing rD with rD/S. However, simulation shows that this system behaves in the 'diminishing returns' manner. This result suggests that enzyme diffusion is not a viable explanation for the 'economy of scale' observation.

Recent work has shown other factors influencing splicing efficiency. For instance, Bentley et al.[7] showed that polymerase elongation speed is involved in splicing regulation: the faster the polymerase speed, the lower the splicing efficiency. To achieve the observed 'economy of scale', polymerase speed should slow down at high transcription levels. However, this assumption does not agree with previous studies[7], indicating polymerase elongation speed is not a possible mechanism for 'economy of scale' behavior. In addition, Luco et al.[8] have recently found that epigenetics also influences splicing regulation. However, the dynamics of epigenetic changes are normally on the order of days[9], not consistent with our experimental timeframe (a few hours). Thus, we can rule out the effects of epigenetics as well.

The other concern is from our experimental design. Our method measured the numbers of the different transcripts at the TAS and used them to calculate splicing efficiency. However, different transcripts could have different residence times at the TAS: the spliced mature RNA can only release from the TAS after coupling 5' capping and polyadenylation[10,11], while spliced introns could diffuse out much faster. This difference influences the number of transcripts at the TAS and could thus impact the splicing efficiency measurement. To investigate this issue, we modulated the parameters related to residence time (g_u and g_m) in our model. Specifically, when scanning g_u from 0.01 to 10 (i.e. 4 orders of magnitude), the system remains in the 'economy of scale' pattern (Fig. S9c). However, when g_u and/or g_m depend on the total amount of transcripts (S + m + u + ES) at the TAS, the 'economy of scale' behavior occurs. Although our system cannot measure different residence times of transcripts at the TAS, we addressed this issue (i.e. differences in residence times) indirectly. For two cells with comparable TAS, as shown in Figure S8, we observed many single transcripts diffusing out from one TAS (bottom cell) but no obvious transcripts from the other (top cell). This diffusion pattern could reflect the residence time of transcripts to some extent. However, we do not have a convincing model to convert the diffusion pattern to residence time. To rule out the effects of different residence times, we sought to quantify only the TASs without single transcripts spreading (as discussed in "TAS classification"). Additionally, to ensure that the economy of scale phenomenon is robust to the effects of different residence time, we quantified the ratio between spliced and unspliced isoforms of RG6 across various transcription levels by qPCR (Fig. S9). This independent measurement, which includes isoforms outside the TAS, thus minimizing the contribution of different residence time, produced a similar 'economy of scale' behavior.

References:

1. Mueller, F. *et al.* FISH-quant: automatic counting of transcripts in 3D FISH images. *Nat. Methods* **10**, 277–278 (2013).
2. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
3. Stetson, P. B. DAOPHOT: A computer program for crowded-field stellar photometry. *Publ. Astro. Soc. Pac.* (1987).
4. Park, T. *et al.* Bayesian estimation of hardness ratios: Modeling and computations. *Astrophys. J.* **652**, 610 (2006).
5. Coath, C. D., Steele, R. C. J. & Fred Lunnon, W. Statistical bias in isotope ratios. *J. Anal. At. Spectrom.* **28**, 52–58 (2013).
6. Ni, J. Z. *et al.* Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* **21**, 708–718 (2007).
7. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
8. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144**, 16–26 (2011).
9. Bintu, L. *et al.* Dynamics of epigenetic regulation at the single-cell level. *Science* **351**, 720–724 (2016).
10. Colgan, D. F. & Manley, J. L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).
11. Shatkin, A. J. & Manley, J. L. The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.* **7**, 838–842 (2000).